（様式７）

学 位 論 文 審 査 結 果 の 要 旨

氏　　名	鯨井俊宏
審 査 委 員	委 員 長　横田　孝義　印 委　　員　木村　周平　印 委　　員　大木　誠　印 委　　員　　印 委　　員　　印
論 文 題 目	Greedy Action Selection and Pessimistic Q-value Updating in Multi-Agent Reinforcement Learning with Sparse Interaction (スパースな干渉下での強化学習におけるグリーディな行動選択と悲観的な Q 値の更新)

審 査 結 果 の 要 旨

機械学習の分野において,マルチエージェント強化学習(MARL)は,複数のロボットがあるタスクを遂行する上で協調作業を行うポリシーを学習する方法などとして有効であると考えられている.しかし,各エージェントの状態と各エージェントがとる行動の組み合わせはエージェント数の増加とともに指数関数的に膨大になり,現実的な問題に適用することは困難となる.これに対し,エージェント間の相互作用が疎な関係にある,すなわち sparse な相互作用を仮定すれば,上記の状態と行動の組み合わせ数、すなわち状態行動空間を飛躍的に縮小することが可能である. 本論文では,エージェント間の疎な関係に着目して従来手法である CQ-learning 法に基づいた 5 種の強化学習方法を提案し,それらの学習アルゴリズムが従来手法に比べて学習能力が優れていることを 5 種類の迷路および 7 種の追跡ゲーム問題による比較で示している.

1 章では本研究の基本になる単一エージェントシステムとマルチエージェントシステムについて説明し,疎なマルチエージェントシステムについて説明している.

2 章では単一エージェントシステムにおける強化学習の原理をレビューし,エージェントの状態遷移が MDP(マルコフ決定過程)に従うと仮定した場合の強化学習方法として 3 種の基本アルゴリズム(すなわち,動的計画法,モンテカルロ法,ＴＤ学習（時間的差分学習）について述べ,特に TD 学習の中で主力である Q ラーニングについて説明している.

3 章ではマルチエージェントシステムにおいて,状態と行動の組み合わせ数の爆発の問題について迷路を例にとり取り上げている.また,この状態行動空間の爆発により学習が妨げられることを示し,sparse interaction（疎な相互作用）の考え方を導入することによって状態行動空間を大幅に小さくすることが可能であることを示し, sparse interaction が仮定できる場合の問題のクラスとして定式化されている Dec-SIMDP(Decentralized Sparse Interaction MDP)について説明している.　さらに Dec-SIMDP に属する問題に対する解法の１つである CQ-Learning(Coordinating Q-learning)によって他のエージェントの状態を考慮すべきか否かを各エージェントが単一エージェント環境での行動に伴う報酬との変化を用いて適切に判断することによって状態行動空間を有効に縮小することが可能であることを示している.

4 章では上記の CQ-Learning には４つの課題が残っていることを指摘している.それらは,(1)事前学習をどのように行えば良いか,(2) ϵ-greedily による不必要な行動の選択が行われる場合があること.(3)楽観的な Q 値(即時報酬)の更新問題 (4)3 つ以上のエージェントが干渉した場合にどうしたらよいか？ これらの課題を解決するために,(1)単一エージェント環境下での事前学習においては十分に状態行動空間を学習するために ε の値を 0.8 と大きな値にする.(2)不必要な行動の探索を行うのは単一エージェント環境において学んだ拡張されていない状態にある場合に限定する. (3)楽観的な Q 値の更新を避けるためにエージェントが他のエージェントとまだ干渉状態にあるか否かを判断することにした.他のエージェントと干渉状態にある場合は悲観的な Q 値の更新を行う.(4) 3 体以上のエージェントが干渉する場合は干渉中のエージェントから一つを選択して行動を起こすことにした.
この方法を 5 種類の迷路を用いて評価した結果,提案手法（GPCQ-learning と呼ぶ）が大幅に性能向上することを確認した.

5 章では追跡ゲームにおいて上記で提案したて GPCQ-learning アルゴリズムが greedy な行動のためにデッドロック状態に陥る場合があることを指摘し,かつ,干渉状態にある場合と単一エージェント状態にある場合とで報酬に差が生じない場合があり,その場合にデッドロック状態にあることの検出が出来ていないことを指摘した.この問題を解決するためにデッドロックの検出論理を実現する方法および状態を拡張していない場合でも Q 値を更新する方法の２つの方法を考案した.7 種の追跡ゲームによって評価を行った結果,性能が大きく向上した.

これらの研究成果は計測自動制御学会(SICE:The Society of Instrument and Control Engineers)の査読付き英文論文誌 JCMSI(Journal of Control, Measurement, and System Integration)に掲載済みであり,また,SICE2018 と PRICAI2019(Pacific Rim International Conference on Artificial Intelligence) の 2 件の査読付き国際会議論文を発表済みである.本学位請求論文は鯨井俊宏氏が本学博士後期課程に在学中に得た研究成果をまとめたものであり,今後の機械学習技術の工学的応用の分野に大きく寄与し得ることから博士（工学）を授与するにふさわしい論文であると判定する.